Please cite as: Ritschard, G., R. B. Losa and P. Origoni (2013). Validating Tree Descriptions of Women's Labor Participation with Deviance-based Criteria. In J.J. McArdle & G. Ritschard (eds), *Contemporary Issues in Exploratory Data Mining in Behavioral Sciences*, Routeledge, New York, pages 128–149

Validating Tree Descriptions of Women's Labor Participation with Deviance-based Criteria

Gilbert Ritschard Fabio B. Losa Pau Origoni

Abstract

This chapter presents a full scaled application of induction trees for non-classificatory purposes. The grown trees are used for highlighting regional differences in the women's labor participation, by using data from the Swiss Population Census. Hence, the focus is on their descriptive rather than predictive power. A first tree provides evidence for three separate analyses for non-mothers, married or widowed mothers, and divorced or single mothers. For each group, trees grown by language regions exhibit fundamental cultural differences supporting the hypothesis of cultural models in female participation.

From the methodological standpoint, the main difficulties with such a non-classificatory use of trees have to do with their validation, since the classical classification error rate does not make sense in this setting. We comment on this aspect and propose deviance-based solutions that are both consistent with our non-classificatory usage and easy to compute.

1 Introduction

Induced decision trees have become, since Breiman et al. (1984), popular multivariate tools for predicting continuous dependent variables and for classifying categorical ones from a set of predictors. They are called *regression trees* when the outcome is quantitative and *classification trees* when it is categorical. Though their primary purpose is to predict and to classify, trees can be used for many other relevant purposes: as exploratory methods for partitioning and identifying local structures in datasets, as well as alternatives to statistical descriptive methods like linear or logistic regression, discriminant analysis, and other mathematical modeling approaches (Fabbris, 1997).

This contribution demonstrates such a *non-classificatory* use of classification trees by presenting a full scaled application (Losa et al., 2006) on female labor market data from the Swiss 2000 Population Census (SPC). The use of trees for our analysis was dictated by our primary interest in discovering the interactions effects of predictors of the women's labor participation. Since

the goal is no longer to extract classification rules, but to understand — from a cross-cultural perspective — the forces that drive women's participation behavior, misclassification rates do not make sense when they are used to validate the trees. We therefore rely on best suited alternative fit criteria that we initially introduced in Ritschard and Zighed (2003) and Ritschard (2006). Our experiment brings insight into the limits and practicability of these criteria for large scale applications.

Apart from these methodological aspects, the practical experiment discussed in this paper is original in at least two respects: 1) the use of trees for microeconomic analysis, which does not appear to be a common domain of application; 2) the use of induction trees for a complete population census dataset.

Section 2 briefly recalls the principle of classification trees. In Section 3, we present the socio-economic research objectives and discuss the main findings. Section 4 is devoted to the validation issue and in Section 5 we apply the introduced deviance-based measures to the trees grown for studying women-participation. Finally, we conclude in Section 6 with an overall evaluation of the experience and of the application of classification trees for non-classificatory purposes.

2 Classification trees principle

Classification trees are grown by seeking, through successive splits of the learning data set, some optimal partition of the predictor space for predicting the outcome class. Each split is done according to the values of one predictor. The process is greedy. At the first step, it tries all predictors to find the "best" split. Then, the process is repeated at each new node until some stopping rule is reached. This requires a local criterion to determine the "best" split at each node. The choice of the criterion is the main difference between the various tree growing methods that have been proposed in the literature, of which CHAID (Kass, 1980), CART (Breiman et al., 1984), C4.5 (Quinlan, 1993) and party (Hothorn et al., 2006) are perhaps the most popular. For our application, we used CART that builds only binary trees by choosing at each step the split that maximizes the gain in purity measured by the Gini index. CART uses relatively loose stopping rules, but proceeds to a pruning round after the preliminary growing phase.

One of the striking features of induction trees is their ability to provide results in a visual form that provides straightforward interpretations. This visual feature, when compared with the outcome of regression models for instance, has exceptional advantages in terms of user-friendliness and in supporting the knowledge discovery process. Furthermore, by their very nature, trees provide a unique description of the predictor interaction effects on the response variable. These advantages remain true as long as the tree does

not become too complex. That is why we chose CART for our analysis, despite the gain in purity seems less appropriate for a non-classificatory purpose than, for example, the strength of association criterion used by CHAID. Indeed, the great readability of the binary CART trees was decisive when compared with the n-ary CHAID trees that had, even at the first level, a much too high number of nodes to allow for any useful interpretation.

As with other statistical modeling approaches, it is essential to assess the quality of the obtained tree before drawing any conclusion from it. Our point is to make it clear that the validation criteria are largely dependant of the pursued goal. Especially, it is worth mentioning that the misclassification rate, which is most often the only validation criterion provided by software programs, is of little help in non-classificatory settings.

3 The applied study

We begin by setting the applied research framework, then we sketch our global analysis procedure and, finally, we present selected findings.

3.1 The topic: Female labor market participation in Switzerland

Female labor market participation reveals significant differences across countries. In Europe, scholars often identify at least two general models: a Mediterranean one (Italy, Greece, Portugal, etc.) versus a model typical

for Central and Northern Europe (Reyneri, 1996). The first is represented by an inverse L-shaped curve of the *activity* or *participation rate* by age, where after a short period of high rate (at entry in the labor market) the proportion of women working or seeking work begins to steadily decline up to retirement. The same graph depicts a M-shaped curve in Central and Northern European countries, characterized by high participation at entry, followed by a temporary decline during the period of motherhood and child-bearing, and a subsequent comeback to work, up to a certain age where the process of definite exit starts.

In this respect, Switzerland is an interesting case. Firstly, Switzerland is a country placed in a nutshell across the Alps, which are considered as one of the cleavages dividing Southern Europe from Central and Northern Europe. Secondly, there are three main languages, spoken by people living in three geographically distinct regions: French in the western part on the border with France, German in the northern and eastern parts on the border with Germany and Austria, and Italian south of the Alps in a region leading to Italy.

The existence of three regions, with highly distinctive historical, social and cultural backgrounds and characters, and the fact that the Italian-speaking part is divided from the other two by the Alps highlight the very specific particularity of this country for a cross-cultural analysis of the female participation in the labor market. Moreover, the fact that the comparative analysis is performed amongst regions of the same country guarantees, de-

spite differences stemming from the Swiss federal system, a higher degree of comparability on a large series of institutional, political and other factors than one would get with cross-country studies.

The idea of the research project was to verify the existence of differing cultural models of female labor market participation, by analysing activity rates and hours worked per week — in terms of proportions of full-timers and part-timers — across the three linguistic regions in Switzerland, by using the SPC 2000 data.

To shortly describe the data, we can say that the Federal Statistical Office made us available a clean census dataset covering the about 7 millions inhabitants of Switzerland. For our study, only the about 3.5 millions women were indeed of interest. In the preprocessing step we disregarded young ($< 20, 23\%$) and elderly ($> 61, 18\%$) women, as well as non Swiss women not born in Switzerland (1.6%), i.e. about 43% of the women. This left us with about 2 millions cases. Finally, we dropped about 350000 cases with missing values, and hence included 1667494 cases into the analysis.

3.2 The empirical research design

The research procedure used classification trees at two different stages, with differing but complementary purposes.

A tree was first grown in what we refer to as the *preliminary step*. Its main goal was, in the spirit of structured induction (Shapiro, 1987) and local pattern detection (Hand, 2002; Rüping, 2005), to find a sound partition of

the analysed population into a limited number of homogeneous groups — homogeneous female labor supply behavior in terms of activity and choice between full-time and part-time employment — over which a tailored analysis could be performed. In other words, in order to avoid an "average" analysis at global level, classification trees have been used to structure the research and to identify those groups of the population which could be used to guide subsequent analysis.

This first step was run on the whole Swiss female population of age 20 to 61, using their *labor market status*[1] as outcome variable, and general socio-demographic characteristics (civil status, mother/non mother, ...) as predictive attributes. From this, a robust partition in three groups was chosen, as the best compromise between level of details for the subsequent analysis and population size of each group. The three groups are the *non-mothers*, the *married or widowed mothers*, and the *divorced or single mothers*. The first group is composed by 609,861 women (36.6%), the second one by 903,527 (54.2%) and the third one by 154,106 (9.2%).

The second application of classification trees took place in the analysis of cross-cultural female labor supply behavior for each selected group. Here again the outcome variable was the *labor market status* of the women. A much broader series of predictive variables was retained however: age, profes-

[1]Labor market status is a categorical variable with four values: full-time active (at least 90% of standard hours worked per week), long part-time active (50% to 90%), short part-time active (less than 50%) and non active, where active means working or seeking for a job.

sion, educational level, mother/non mother, number of kids, age of last-born kid, type of household, etc. Before growing trees, we carried out a series of simple bivariate analyses between the labor market status and each selected predictive attribute. This helped to identify the most relevant attributes for the retained cross-cultural perspective. The analysis of their raw impact on the labor status provided useful indications on how important each one is when it comes to explaining the female labor supply behavior.

Classification trees have been produced separately for each region and then compared, as described in the next section, in order to analyse cultural patterns in the participation behavior of the main language regions in Switzerland. At this stage, classification trees and traditional analyses have been used in a complementary way allowing for interplay between them. This proved to be highly productive in stimulating the knowledge discovery process as well as in analysis and understanding of relevant phenomena.

It is worth mentioning here that the final trees retained are simplified versions of those that resulted from the stopping and pruning criteria. They were selected on the basis of comprehensibility and stability factors. We checked for instance that the splits retained stayed the same when removing randomly 5% of the cases from the learning data set.

3.3 Results

3.3.1 Definition of the groups of analysis.

The three groups identified in the preliminary local pattern detection step appear to exhibit a high degree of inter-group diversity combined with a significant intra-group homogeneity. Inter-group diversity is highlighted by the very specific participation rates by age depicted in Figure 1.[2]

[Figure 1 about here.]

Comparison of part-time versus full-time employment reinforces the picture by highlighting the very different choices made by working women of the three groups: a majority of the non-mothers choose full-time employment all along their professional life, divorced and single mothers switch from part-time jobs during motherhood and early childbearing to full-time (or long part-time) jobs, and the married and widowed mothers prefer short part-time employment in the majority of cases.

3.3.2 The determinants of labor supply behavior of divorced and single mothers.

In order to identify cultural models of female labor supply, three trees (one per region) were generated for each group. These — in combination with the results of the traditional bivariate analyses — were compared and thoroughly

[2]Figure 1 demonstrates that the M- or L-shaped curves encountered in cross-country studies may result from the superposition of group specific curves.

analyzed in terms of structure and results. We give hereafter a very brief overview of the main results for the third group, i.e. divorced or single mothers.[3] For details interested readers may consult the research report (Losa and Origoni, 2005). In Figure 2 and 3, white background is used for nodes with a majority of non active women, light grey for a majority of part-timers and dark grey for a majority of full-timers. We see that opting for inactivity seems to be much more frequent in the Italian speaking region.

[Figure 2 about here.]

Profession and *age of the mother* point out specific groups with particular distinct behaviors. The former puts apart a group of professions — in the fields of health, education, sciences, etc. — which are known to be characterized by high proportions of part-time jobs. Age plays a central role in the Swiss Italian (Figure 2) and in the (not shown) Swiss German tree by clearly splitting the period of active life (up to age 54-55), from that of the definite withdrawal from the labor market.

[Figure 3 about here.]

The *age of the last-born child* appears as the most discriminative factor in all the regions demonstrating the very central role within the family-work conflict of being mother for the women of this group, who live mainly in single-parent households. The most significant differences across the Swiss

[3]For space reasons, only the (slightly simplified) trees of the Italian and French speaking regions are presented.

language regions appear in this variable, namely on its position in the tree, its split values and the distribution in the classes of the resulting partition. There is a high proportion of inactivity among Swiss Italian women living in single-parent household when last-born child is 2 years old or younger. This proportion decreases for the first time when the child is 3 (access to public kindergarten) and for the second time when the child reaches 6 (access to primary school). Swiss German women also quit the labor market, but re-enter sooner, while Swiss French are almost indifferent to this factor, showing constant activity rates per age of last-born child.

In all the three regions, *educational level* has a strong influence on female labor supply. The higher the educational level, the higher the proportion of active mothers and the lower the proportion of full-timers. This double effect is particularly evident in the Italian speaking and German speaking regions, when last-born child is very young (less than 4 respectively 6 years). Mothers with elementary or intermediate level education decide in the majority of cases to quit their jobs and to stay at home during this period, while mothers of higher education work on a part-time basis.

The *presence of the partner* and the *number of children*, which strongly influence the behavior of married women have only limited effect on divorced women.

4 Validating the tree descriptive ability

For the reliability of the description, individual predictions do not matter. Rather, we focus on the posterior distribution of the response variable, i.e., on the distribution conditioned by the values of the predictors. These posterior distributions are the columns of the target table. Our concern is thus to measure how well a tree may predict this target table. This is a goodness-of-fit issue very similar to that encountered in the statistical modeling of multiway cross tables. According to our knowledge, however, it has not been addressed so far for induced trees. Textbooks, like Han and Kamber (2006), Hand et al. (2001) for example, do not mention it, and, as far as this model assessment issue is concerned, statistical learning focuses almost exclusively on the statistical properties of the classification error rate (see for example Berk 2009 or Hastie et al. 2001 chap. 7).

In statistical modeling, e.g. linear regression, logistic regression or more generally generalized linear models (GLM), the goodness-of-fit is usually assessed by two kinds of measures. On the one hand, indicators such as the coefficient of determination R^2 or pseudo R^2's tell us how better the model does than some naive baseline model. On the other hand we measure, usually with divergence or deviance statistics, how well the model reproduces some target or, in other words, how far we are from the target.

Our contribution is a trick that permits to use this statistical machinery with induced trees. The trick allows us to propose, among others, an adapted

form of the Likelihood Ratio deviance statistic with which we can test statistically the significance of any expansion of a tree. Other criteria discussed are R^2 like measures and the powerful model selection AIC and BIC criteria.

Before describing the deviance, we start by introducing an illustrative example data set that will serve all along the section. We then specify the notations, terminology and concepts that we use.

4.1 Illustrative example

We consider a fictional example where we are interested in predicting the civil status (married, single, divorced/widowed) of individuals from their gender (male, female) and sector of activity (primary, secondary, tertiary). The civil status is the outcome or response variable, while gender and activity sector are the predictors. The data set is composed of the 273 cases described in Table 1.

[Table 1 about here.]

4.2 Terminology and notations

Classification trees are grown by seeking, through recursive splits of the learning data set, some optimal partition of the predictor space for predicting the outcome class. Figure 4 shows the tree grown for our illustrative data.

[Figure 4 about here.]

A *leaf* is a terminal node. There are 4 leaves in Figure 4.

In the machine learning community, predictors are also called attributes and the outcome variable the predicted attribute. The values of the outcome variable are called the classes. We prefer using "outcome values" to avoid confusion with the classes of the population partition defined by the leaves.

We call *profile* a vector of predictor values. For instance, (female, tertiary) is a profile in Table 1.

We call *target table* and denote by T the contingency table that cross classifies the outcome values with the set of possible profiles. As shown in Table 2, there are 6 possible profiles for our data.

[Table 2 about here.]

Notice that the root node contains just the marginal distribution of the outcome variable. It is useful also to point out that the columns of the target table are just the leaves of a maximally developed tree (see the right side in Figure 5). We call *saturated tree* this maximally developed tree.

The count in cell (i, j) of the target table T is denoted n_{ij}. We designate by $n_{.j}$ and $n_{i.}$ the total of respectively the jth column and ith row.

4.3 The deviance

Having defined the target table, we propose using the deviance for measuring how far the induced tree is from this target (Figure 5). By comparing with the deviance between the root node and the target, we should also be able to

evaluate the overall contribution of the predictors, i.e. what is gained over not using any predictor.

The general idea of the deviance of a statistical model m is to measure how far the model is from the target, or more specifically how far the values predicted by the model are from the target. In general (see for instance McCullagh and Nelder, 1989), this is measured by minus twice the log-likelihood of the model ($-2\text{LogLik}(m)$) and is just the log-likelihood ratio Chi-square in the modeling of multiway contingency tables (Agresti, 1990). For a 2 way $r \times c$ table, it reads for instance

$$D(m) \;=\; 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\,\ln\left(\frac{n_{ij}}{\hat{n}_{ij}}\right)\;, \tag{1}$$

where $\hat{n}_{ij}$ is the estimation of the expected count provided by the model for cell (i,j). The likelihood is obtained assuming simply a multinomial distribution which is by noway restrictive. Under some regularity conditions (see for instance Bishop et al., 1975, chap. 4), the Log-Likelihood Ratio statistic has an approximate Chi-square distribution when the model is correct. The degrees of freedom d are given by the difference between the number of cells and the number of free parameters of the model.

The advantage of the deviance over for instance the Pearson Chi-square is an additivity property that permits to test the difference between a model m_1 and a restricted version m_2 with the difference $D(m_2|m_1) = D(m_2) - D(m_1)$. This difference has indeed also an approximate Chi-square distribution when

the restricted model is correct. Its number of degrees of freedom equals the difference $d_2 - d_1$ in degrees of freedom for each model.

[Figure 5 about here.]

4.4 Deviance for a tree

We have already defined the target table for a classification tree with discrete attributes. Hence, we should be able to compute a deviance for the tree. We face two problems however:

1. How do we compute the predicted counts $\hat{n}_{ij}$ from the induced tree ?

2. What are the degrees of freedom ?

To answer these questions we postulate a (non restrictive) multinomial distribution of the outcome variable for each profile. More specifically, we assume a discrete distribution

$$\vec{p}_j = (p_{1|j}, \ldots, p_{r|j}) \ ,$$

where $p_{i|j}$ is the probability to be in state i of the outcome variable for a case with profile $\vec{x}_j$.

A tree with $q \leq c$ leaves can be seen as a model of the target table. It states that the probability $p_{i|j}$ of being in the ith value of the outcome

variable is equal for all profiles j belonging to a same leaf k, i.e.

$$p_{i|j} = p^*_{i|k}, \quad \text{for all } \vec{x}_j \in \mathcal{X}_k, k = 1, \ldots, q \ ,$$

where $\mathcal{X}_k$ stands for the set of profiles of leaf k. The tree parameterizes the rc probabilities $p_{i|j}$ in terms of rq parameters $p^*_{i|k}$, which leaves

$$d = (r-1)(c-q) \text{ degrees of freedom } . \tag{2}$$

The probabilities $p^*_{i|k}$'s are estimated by the observed proportions, i.e $\hat{p}^*_{i|k} = n_{ij}/n_{\cdot j}$. Estimates of the probabilities $p_{i|j}$ are derived from those of the $p^*_{i|k}$'s, i.e. $\hat{p}_{i|j} = \hat{p}^*_{i|k}$ when $\vec{x}_j \in \mathcal{X}_k$.

For given n_j's and given distributions $\vec{p}_j$, the expected counts for a profile $\vec{x}_j$ is $n_{\cdot j} p_{i|j}$, for $i = 1, \ldots, r$. Now, replacing the $p_{i|j}$'s by their estimates, we get estimates $\hat{n}_{ij}$ of the expected counts:

$$\hat{n}_{ij} \quad = \quad n_{\cdot j} \hat{p}^*_{i|k} \quad \text{for all } \vec{x}_j \in \mathcal{X}_k, k = 1, \ldots, q \ . \tag{3}$$

Table 3 shows the counts predicted this way from the tree in Figure 4.

[Table 3 about here.]

Considering the counts of the target table and the estimates (3), the deviance $D(m)$ of a tree m can be computed using formula (1). For our example we find $D(m) = 1.69$. The number of degrees of freedom is $d(m) = (3-1)(6-4) = 4$. The obtained deviance being much less than $d(m)$, it is

clearly not statistically significant indicating that the induced tree fits well the target T.

4.5 Using the deviance

The approximated Chi-square distribution of the deviance holds when the expected counts per cell are all say greater than 5. This is rarely the case when the number of predictors is large. Hence, the deviance will not be so useful for testing the goodness-of-fit. Note that we have exactly the same problem with, for instance, logistic regression.

Nevertheless, the difference in the deviance for two nested trees will have a Chi-square distribution, even when the deviances themselves do not.

$$D(m_2|m_1) = D(m_2) - D(m_1) \sim \chi^2 \text{ with } d_2 - d_1 \text{ degrees of freedom} \quad .$$

Thus, the main interest of the deviance is to test differences between nested trees. A special case is testing the difference with the root node with $D(m_0|m)$, which is the equivalent of the usual Likelihood Ratio Chi-square statistic used in logistic regression.

For our example, we have $D(m_0|m) = 167.77$ for 6 degrees of freedom. This is clearly significant and demonstrates that the tree describes the outcome significantly better than independence (root node). The predictors bring significant information.

As a further illustration, let us test if pruning the branches below "female"

in the tree of Figure 4 implies a significant change. The reduced tree m_1 has a deviance $D(m_1) = 32.4$ for 6 degrees of freedom. This is statistically significant, indicating that the reduced tree does not fit the target correctly. The difference with the induced tree m is $D(m_1|m) = 32.4 - 1.7 = 30.7$ for 2 degrees of freedom. This is also significant and demonstrates that pruning the branch deteriorates significantly the deviance.

4.6 Deviance-based quality measures

It is very convenient to measure the gain in information in relative terms. Pseudo R^2's, for instance, represent the proportion of reduction in the root node deviance that can be achieved with the tree. Such pseudo R^2's come in different flavors. McFadden (1974) proposed simply $\big(D(m_0) - D(m)\big)/D(m_0)$. A better choice is the improvement of Cox and Snell (1989)'s proposition suggested by Nagelkerke (1991):

$$R^2_{\text{Nagelkerke}} = \frac{1 - \exp\{\frac{2}{n}\big(D(m_0) - D(m)\big)\}}{1 - \exp\{\frac{2}{n}D(m_0)\}} \ .$$

The McFadden pseudo R^2 is 0.99, and with Nagelkerke formula we get 0.98.

We may also consider the percent reduction in uncertainty of the outcome distribution for the tree as compared with the root node. The uncertainty coefficient u of Theil (1970), which reads $u = D(m_0|m)/(-2\sum_i n_{i\cdot} \ln(n_{i\cdot}/n))$ in terms of the deviance, and the association measure τ of Goodman and Kruskal (1954) are two such measures. The first is the proportion of reduc-

tion in Shannon's entropy and the second in quadratic entropy. These two indexes produce generally very close values. They evolve almost in a quadratic way from no association to perfect association (Olszak and Ritschard, 1995). Their square root is therefore more representative of the position between these two extreme situations. For our induced tree, we have $\sqrt{u} = 0.56$, and $\sqrt{\tau} = 0.60$, indicating that we are a bit more than half way to full association. For the reduced tree m_1 (pruning branch below female), these values are smaller $\sqrt{u} = 0.51$, and $\sqrt{\tau} = 0.57$ indicating that the pruned branch bears some useful information about the distribution.

From the deviance, we can derive AIC and BIC information criteria. For instance, the BIC value for a tree m is

$$\text{BIC}(m) = D(m) - d\ \ln(n) + \text{constant}\ \ ,$$

where n is the number of cases and d the degrees of freedom in the tree m. The constant is arbitrary, which means that only differences in BIC values matter. Recall that following Raftery's rules of thumb (Raftery, 1995), a difference in BIC values greater than 10 provides strong evidence for the superiority of the model with the smaller BIC in terms of trade-off between fit and complexity.

4.7 Computational aspects

Though the deviance could easily be obtained on our simple example, its practical use on real life data raises two major issues.

1. Existing softwares for growing trees do not provide the deviance. Furthermore, most of them do not provide the data needed to compute the target table and the estimates $\hat{p}_{i|j}$ in an easily usable form .

2. The number of possible distinct profiles which defines the number c of columns of the target table rapidly becomes excessively large when the number of predictors increases. Theoretically, denoting by c_v the number of values of the variable $x_v, v = 1, \ldots, V$, the number of distinct profiles may be as large as $\prod_v c_v$, which may become untractable.

Regarding the *first point*, we need to compute the "profile" variable, i.e., assign to each case a profile value. The profile variable can be seen as a composite variable x_{prof} with a unique value for each cell of the cross classification of all predictors x_v. Assuming that each variable has less than 10 values, we can compute it, for example, by using successive powers of 10

$$x_{prof} = \prod_{v=1}^{V} 10^{v-1} x_v \ .$$

We need also a "leaf" variable x_{leaf} that indicates to which leaf each case belongs. Here we have to rely on tree growing softwares that either directly produce this variable (`rpart`, Therneau and Atkinson 1997, or `party`,

Hothorn et al. 2006), or like AnswerTree (SPSS, 2001) for instance, generate rules for assigning the leaf number to each case.

The next step is to compute the counts of the target table and those of the leaf table resulting from the cross tabulation of the outcome variable with the leaf variable. This can be done by resorting to softwares that directly produce cross tables. However, since the number of columns, especially that of the columns of the target table, may be quite large and the tables very scarce, a more careful coding that would take advantage of the scarcity is a real concern. A solution is to aggregate cases by profiles and outcome values, which is for instance easily done with software such as SPSS. Creating a similar file by aggregating by leaves and outcome values, the resulting files can then be merged together so as to assign the leaf data to each profile. From here, it is straightforward to get the estimated counts with formula (3) and then compute the deviance $D(m)$ with formula (1). Figure 6 shows the SPSS syntax we used for getting the deviance of our example induced tree.

[Figure 6 about here.]

An alternative solution that can be used by those who do not want to write code, is to use the Likelihood Ratio Chi-square statistic that most statistical packages provide for testing the row-column independence in a contingency table. For the target table this statistic is indeed the deviance $D(m_0)$ between the root node m_0 and the target, while for the leaf table it is the deviance $D(m_0|m)$ between the root node and the leaf table associated

to the induced tree. The deviance for the model is then just the difference between the two (see Figure 5)

$$D(m) \;=\; D(m_0) - D(m_0 \mid m) \;\; .$$

For our example, we obtain with SPSS $D(m_0) = 169.46$ and $D(m_0|m) = 167.77$, from which we deduce $D(m) = 169.46 - 167.77 = 1.69$. This is indeed the value we obtained by applying directly formula (1). Note that this approach is limited by the maximal number of columns (or rows) accepted for cross tables. This is for instance 1000 in SPSS 13, which makes this approach unapplicable when the number of possible profiles exceeds this number.

Let us now turn to the *second issue*, i.e. the possibly excessive number of a priori profiles. The solution we propose is to consider partial deviances. The idea is to define the target table from the mere predictors retained during the growing process. This will reduce the number of variables. We could go even further and group the values of each predictors according to the splits used in the tree. For instance, if the induced tree leads to the 3 leaves "male", "female and primary sector", "female and non primary sector", we would not distinguish between secondary and tertiary sectors. There would thus be 4 profiles — instead of 6 — for the target table, namely "male and primary sector", "male and non primary sector", "female and primary sector", "female and non primary sector".

The resulting target table T^* is clearly somewhat arbitrary. The consequence is that the partial deviance, i.e. the deviance $D(m|m_{T^*})$ between the tree m and T^*, has no real meaning by itself. However, we have $D(m) = D(m|m_{T^*}) + D(m_{T^*})$ thanks to the additivity property of the deviance. It follows that $D(m_2) - D(m_1) = D(m_2|m_{T^*}) - D(m_1|m_{T^*})$. The difference in the partial deviance of two nested trees m_1 and m_2 remains unchanged, whatever target m_{T^*} is used. Thus, all tests based on the comparison of deviances, between the fitted tree and the root node for example, remain applicable.

The partial deviance can also be used for defining AIC and BIC criteria, since only differences in the values of the latter matter. Pseudo R^2's, however, are not very informative when computed from partial deviances, due to the arbitrariness of the target table. It is preferable to consider the percent reduction in uncertainty, which does not depend on the target table, and to look at the square root of Theil's u or Goodman and Kruskal's τ.

5 Validating the women's participation trees

We must first define the target table in order to compute the deviance for our three regional trees. As explained above, this is quite easy as long as only a limited number of attributes with each a limited number of values are used. For our real full-scale application, it happened, nevertheless, to be a virtually unmanageable task. Indeed, cross tabulating the observed values of

the attributes considered gives rise to more than a million different profiles, i.e., columns for the target table.

We therefore considered only a partial deviance $D(m|m_{T^*})$ that measures the departure from the partition m_{T^*} defined by the mere split values used in the tree. In other words, we compare the partition defined by the tree with the finest partition that can be achieved by combining the groups of values defined by the splits.

For our application, we obtained the partial deviances with SPSS. Two deviances were computed, namely $D(m_0|m_{T^*})$ and $D(m_0|m)$, where m_0 is the root node and m the fitted tree. We first recoded the attributes so as to group the values that remain together all over the tree. It was then easy to build a profile variable taking a different value for each observed combination of the recoded values. The target table m_{T^*} results from the cross tabulation of this profile variable with the outcome variable, i.e., the type of participation in the labor market.

The deviance $D(m_0|m_{T^*})$ is finally simply the independence Log Likelihood Ratio Chi-square statistic (LR) for this target table. Likewise, the deviance $D(m_0|m)$ between the root node and the fitted tree is the LR statistic for the table that cross tabulates the leave number with the response variable. Since the trees were grown with Answer Tree (SPSS, 2001), we readily obtained the leave number of each case with the SPSS code generated by this software. The deviance $D(m|m_{T^*})$ that measures how far the tree is from the

target, is obtained as the difference between those two computed deviances:

$$D(m|m_{T^*}) = D(m_0|m_{T^*}) - D(m_0|m) \ .$$

Similar relations hold for the degrees of freedom. Recall, however, that the partial deviance has no real meaning by itself. Its interest lies in that it permits to testing statistically differences between nested trees.

We also derive BIC values from the partial deviance. This is not restrictive since only differences in the values of the latter matter. We thus compute the BIC value for a tree m as

$$\text{BIC}(m) = D(m|m_{T^*}) - \ln(n)(c^* - q)(\ell - 1) \ ,$$

where n is the number of cases, c^* is the number of different profiles in the target table m_{T^*}, q the number of leaves of the tree and ℓ the number of outcome classes, i.e., in our application, the four types of participation in the labor market. The product $(c^* - q)(\ell - 1)$ gives the degrees of freedom associated with the partial deviance.

It is also very convenient to measure the gain in information in relative terms. Pseudo R^2's are not very informative when computed from partial deviances, due to the arbitrariness of the target table. It is preferable to consider the percent reduction in uncertainty about the outcome distribution achieved with the tree when compared to the root node such as measured, for instance, by the uncertainty coefficient u of Theil (1970).

[Table 4 about here.]

Table 4 reports some of the quality figures we have computed for each of the three regional trees: CHI for the Italian speaking, CHF for the French speaking and CHG for the German speaking region. The deviances $D(m_0|m)$ are all very large for their degrees of freedom. This tells us that the grown trees clearly improve the description as compared to the root node. The deviances $D(m|m_{T^*})$, not shown here, are also very large indicating that there remains room for improving the fit. The difference ΔBIC in the BIC values between the root node and the grown trees lead to a similar conclusion. They are largely superior to 10, providing evidence of the superiority of the grown trees over the root node. For CHI and CHF, the BIC values of the grown trees are also much smaller than those of the associated saturated trees. This is not the case, however, for CHG. There is thus definite room for improvement in this last case. Remember, however, that we are interested in pointing out the main forces that drive the female participation in the labor market. Hence, we have a comprehension purpose, for which increasing complexity would undoubtedly be counter productive. This is typical in socio-economic modeling, where we cannot let the modeling process be entirely driven by purely statistical criteria. Indeed, the trees need to make sense.

The Theil uncertainty coefficient u seems to exhibit a low proportion of gain in uncertainty. However, looking at its square root, we see that we have covered about 25% of the distance to perfect association. Furthermore, the values obtained should be compared with the maximal values that can be

achieved with the attributes considered. For the target table, which retains a partition into c^* classes, the u is, respectively, .28, .24 and .23. The square root of these values is about .5, i.e. only about twice the values obtained for the trees. Thus, with the grown trees that define a partition into q classes only, we are about half the way from the target table.

To illustrate how these measures can be used for tree comparison, consider the simplified tree in Figure 2 obtained from CHI by pruning a branch grown from the node "age 55-61 years". The original tree CHI has $q = 12$ leaves, while the simplified tree has only 9 terminal nodes. For the latter, we get $D(m_0|m) = 799.4$ with $d = 24$, which leads to a significant difference in deviances of 22.8 for 9 degrees of freedom. The $\Delta(BIC)$ between the two trees is however 55.1 in favor of the simplified tree, the loss in fit being more than compensated by the complexity reduction. This statistically grounds the retained simplification.

6 Conclusion

The experiment reported demonstrates the great potential of classification trees as an analytical tool for investigating socio-economic issues. Especially interesting is the visual tree outcome. For our study, this synthetic view of the relatively complex mechanisms that steer the way women decide about their participation in the labor market provided valuable insight into the studied issue. It allowed us to highlight regional cultural differences in the

interaction effects of attributes like age of last-born child, number of children, profession and education level that would have been hard to uncover through regression analysis, for example.

It is worth mentioning that generating reasonably sized trees is essential when the purpose is to describe and understand underlying phenomenon. This is not the case with classification settings. Indeed, complex trees with many levels and hundred of leaves, even with excellent classification performance in generalization, would be too confusing to be helpful. Furthermore, in a socio-economic framework, like that considered here, the tree should make sense from the social and economic standpoint. The tree outcomes should therefore be confronted with other bivariate analyses and modeling approaches. Our experience benefited a great deal from this interplay.

Now, as end users, we had to face the lack of suitable validation measures provided by the tree growing software programs for our non-classificatory purpose. The main novelty proposed here is the partial deviance and the efficient way to compute it. The relevance of the partial deviance is based on the additivity property of the deviance. Alternative chi-square divergence measures (Pearson for example) could be considered. However, since they do not share the additivity property, we could not as easily derive partial forms of them.

Although we were able to obtain relevant indicators and statistics afterwards by means of classical cross tabulation outcomes, we would urge software developers to include such validation measures in their software output.

Even more, we are convinced that better descriptive trees can be generated when maximal change in overall deviance or BIC values is used as a criterion for growing trees.

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Berk, R. A. (2009). *Statistical Learning from a Regression Perspective*. New York: Springer.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis*. Cambridge MA: MIT Press.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Cox, D. R. and E. J. Snell (1989). *The Analysis of Binary Data* (2nd ed.). London: Chapman and Hall.

Fabbris, L. (1997). *Statistica multivariata: analisi esplorativa dei dati*. Milano: McGraw Hill.

Goodman, L. A. and W. H. Kruskal (1954). Measures of association for cross classifications. *Journal of the American Statistical Association 49*, 732–764.

Han, J. and M. Kamber (2006). *Data Mining: Concept and Techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

Hand, D. J. (2002). The framing of decisions as distinct from the making of decisions. In A. M. Herzberg and R. W. Oldford (Eds.), *Statistics, Science, and Public Policy VI: Science and Responsibility*, Kingston, Ont., pp. 157–161. Queen's University.

Hand, D. J., H. Mannila, and P. Smyth (2001). *Principles of Data Mining*. Adaptive Computation and Machine Learning. Cambridge MA: MIT Press.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics 15*(3), 651–674.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics 29*(2), 119–127.

Losa, F. B. and P. Origoni (2005). The socio-cultural dimension of women's labour force participation choices in Switzerland. *International Labour Review 44*(4), 473–494.

Losa, F. B., P. Origoni, and G. Ritschard (2006). Experiences from a socio-economic application of induction trees. In L. Todorovski, N. Lavrač, and

K. P. Jantke (Eds.), *Discovery Science, 9th International Conference, DS 2006, Barcelona, October 7-10, 2006, Proceedings*, Volume LNAI 4265, pp. 311–315. Berlin Heidelberg: Springer.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. London: Chapman and Hall.

McFadden, D. (1974). The measurment of urban travel demand. *Journal of Public Economics 3*, 303–328.

Nagelkerke, N. J. D. (1991). A note on the general definition of the coefficient of determination. *Biometrika 78*(3), 691–692.

Olszak, M. and G. Ritschard (1995). The behaviour of nominal and ordinal partial association measures. *The Statistician 44*(2), 195–212.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. Marsden (Ed.), *Sociological Methodology*, pp. 111–163. Washington, DC: The American Sociological Association.

Reyneri, E. (1996). *Sociologia del mercato del lavoro*. Bologna: Il Mulino.

Ritschard, G. (2006). Computing and using the deviance with classification trees. In A. Rizzi and M. Vichi (Eds.), *COMPSTAT 2006 - Proceedings in Computational Statistics*, pp. 55–66. Berlin: Springer.

Ritschard, G. and D. A. Zighed (2003). Goodness-of-fit measures for induction trees. In N. Zhong, Z. Ras, S. Tsumo, and E. Suzuki (Eds.), *Foundations of Intelligent Systems, ISMIS03*, Volume LNAI 2871, pp. 57–64. Berlin: Springer.

Rüping, S. (2005). Learning with local models. In K. Morik, J.-F. Boulicaut, and A. Siebes (Eds.), *Local Pattern Detection*, Volume 3539 of *LNCS*, Berlin, pp. 153–170. Springer.

Shapiro, A. D. (1987). *Structured Induction in Expert System*. Wokingham: Adison-Wesley.

SPSS (Ed.) (2001). *Answer Tree 3.0 User's Guide*. Chicago: SPSS Inc.

Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology 76*, 103–154.

Therneau, T. M. and E. J. Atkinson (1997). An introduction to recursive partitioning using the RPART routines. Technical Report Series 61, Mayo Clinic, Section of Statistics, Rochester, Minnesota.

List of Figures

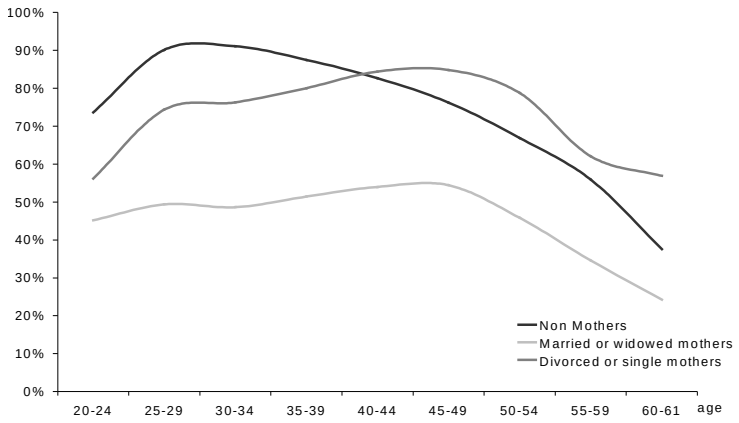

Figure 1: Activity rates by age of the three groups selected

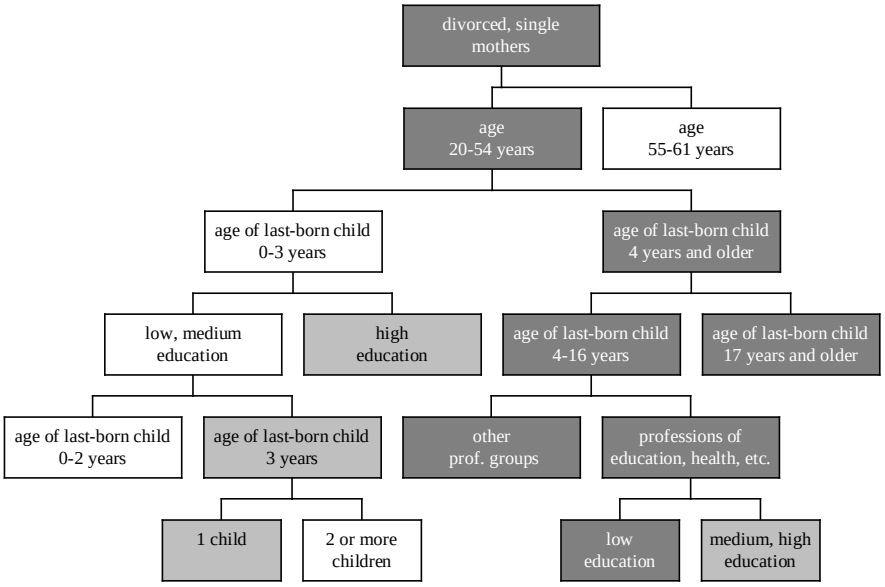

Figure 2: Tree for participation of divorced or single mothers, Italian speaking region

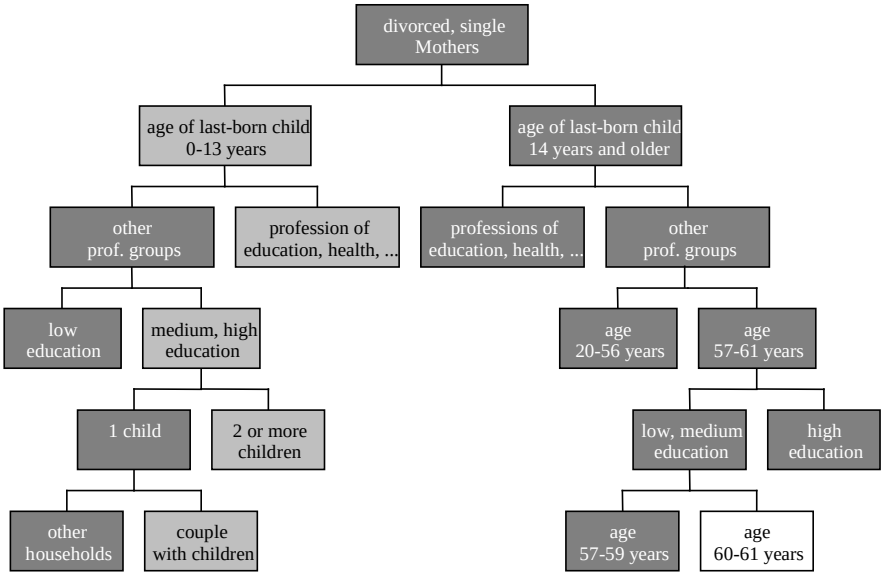

Figure 3: Tree for participation of divorced or single mothers, French speaking region

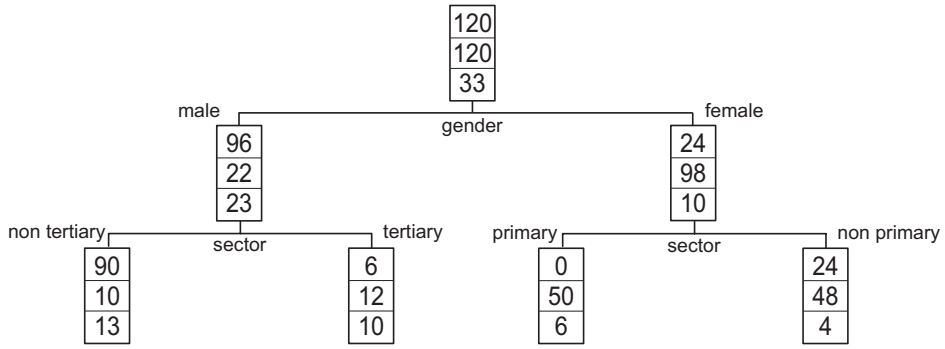

Figure 4: Example: Induced tree for civil status (married, single, divorced/widowed)

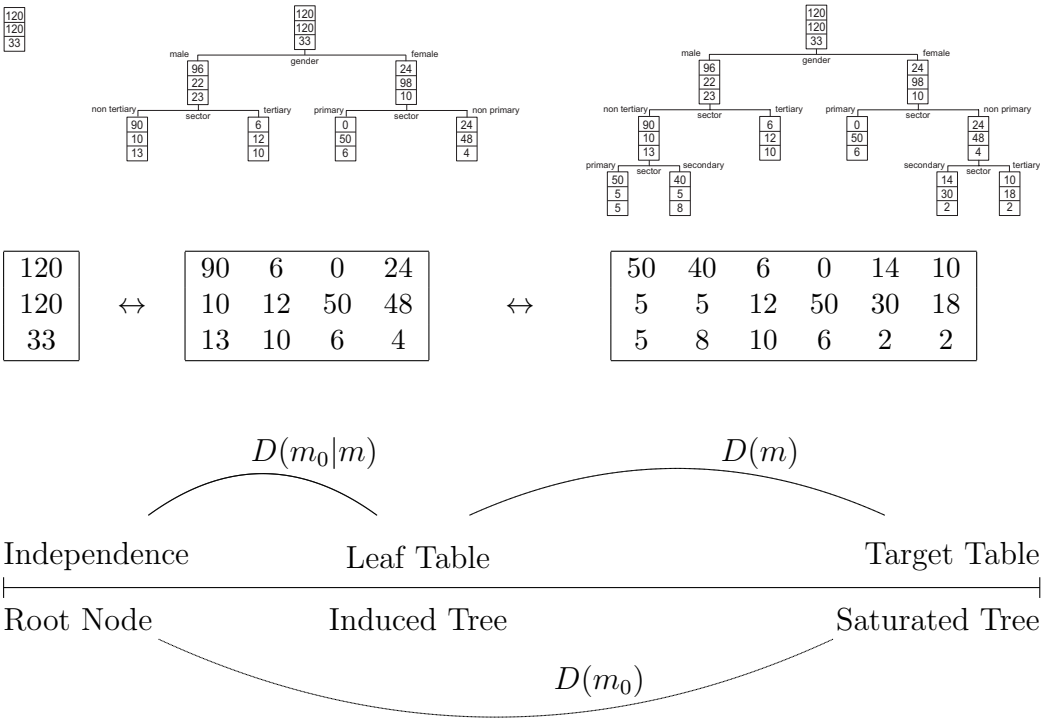

Figure 5: Deviance

```
GET FILE='civst_gend_sector.sav'.
compute profiles
  = ngender*10^1 + nsect.
**Rules generated by AnswerTree**.
IF (ngender NE 2) AND (nsect NE 3)
  leaf = 3.
IF (ngender NE 2) AND (nsect EQ 3)
  leaf = 4.
IF (ngender EQ 2) AND (nsect EQ 1)
  leaf = 5.
IF (ngender EQ 2) AND (nsect NE 1)
  leaf = 6.
END IF.
**Computing the deviance**.
SORT CASES BY profiles .
AGGREGATE
  /OUTFILE='profiles.sav'
  /PRESORTED
  /BREAK=profiles
  /prof_mar = PIN(ncivstat 1 1)
  /prof_sgl = PIN(ncivstat 2 2)
  /prof_div = PIN(ncivstat 3 3)
  /leaf = first(leaf)
  /nj=N.
SORT CASES BY leaf.
AGGREGATE
  /OUTFILE='leaves.sav'
  /PRESORTED
  /BREAK=leaf
  /leaf_mar = PIN(ncivstat 1 1)
  /leaf_sgl = PIN(ncivstat 2 2)
  /leaf_div = PIN(ncivstat 3 3)
  /nj=N.
GET FILE='profiles.sav'.
SORT CASES BY leaf.
MATCH FILES /FILE=*
 /TABLE='leaves.sav'
 /RENAME (nj = d0)
 /DROP d0
 /BY leaf.
COMPUTE pre_mar=leaf_mar*nj/100.
COMPUTE pre_sgl=leaf_sgl*nj/100.
COMPUTE pre_div=leaf_div*nj/100.
COMPUTE n_mar=prof_mar*nj/100.
COMPUTE n_sgl=prof_sgl*nj/100.
COMPUTE n_div=prof_div*nj/100.
**Restructuring data table**.
VARSTOCASES
 /MAKE count
   FROM n_mar n_sgl n_div
 /MAKE pre
   FROM pre_mar pre_sgl pre_div
 /INDEX= Index1(3)
 /KEEP = profiles leaf
 /NULL = DROP
 /COUNT= nclass .

SELECT IF count > 0.
COMPUTE
  deviance=2*count*ln(count/pre).
SORT CASES BY leaf profiles.
COMPUTE newleaf = 1.
IF (leaf=lag(leaf,1))
    newleaf = 0.
COMPUTE newprof = 1.
IF (profiles=lag(profiles,1))
    newprof = 0.
COMPUTE one = 1.
FORMAT one (F2.0)
  /newleaf newprof (F8.0).
**Results in one row table**.
AGGREGATE
  /OUTFILE='deviance.sav'
  /PRESORTED
  /BREAK=one
  /deviance = sum(deviance)
  /nprof    = sum(newprof)
  /nleaves  = sum(newleaf)
  /nclass   = first(nclass)
  /ncells   = N.
GET FILE='deviance.sav'.
**DF and Significance**.
COMPUTE
  df=(nclass-1)*(nprof-nleaves).
COMPUTE
  sig=CDF.CHISQ(deviance,df).
EXECUTE.
```

Figure 6: SPSS syntax for computing the deviance of the tree

List of Tables

Table 1: Example: The data set

Civil status	Gender	Activity sector	Number of cases
married	male	primary	50
married	male	secondary	40
married	male	tertiary	6
married	female	primary	0
married	female	secondary	14
married	female	tertiary	10
single	male	primary	5
single	male	secondary	5
single	male	tertiary	12
single	female	primary	50
single	female	secondary	30
single	female	tertiary	18
divorced/widowed	male	primary	5
divorced/widowed	male	secondary	8
divorced/widowed	male	tertiary	10
divorced/widowed	female	primary	6
divorced/widowed	female	secondary	2
divorced/widowed	female	tertiary	2

Table 2: Target table

	male			female			
	primary	secondary	tertiary	primary	secondary	tertiary	total
married	50	40	6	0	14	10	120
single	5	5	12	50	30	18	120
div./wid.	5	8	10	6	2	2	33
total	60	53	28	56	46	30	273

Table 3: Predicted counts

	male			female			
	primary	secondary	tertiary	primary	secondary	tertiary	total
married	47.8	42.2	6	0	14.5	9.5	120
single	5.3	4.7	12	50	29.1	18.9	120
div./wid.	6.9	6.1	10	6	2.4	1.6	33
total	60	53	28	56	46	30	273

Table 4: *Trees quality measures*

	q	c^*	p	n	$D(m_0\|m)$	d	sig.	ΔBIC	u	$\sqrt{u}$
CHI	12	263	299	5770	822.2	33	.00	536.4	.056	.237
CHF	10	644	674	35239	4293.3	27	.00	4010.7	.052	.227
CHG	11	684	717	99641	16258.6	30	.00	15913.3	.064	.253